Journal of Modern Applied Statistical Methods

Volume 4 | Issue 2 Article 7

11-1-2005

A Discretized Approach to Flexibly Fit Generalized Lambda Distributions to Data

Steve Su
Epi-stat Division, George Institute for International Health, ssu@thegeorgeinstitute.org

Journal of Modern Applied Statistical Methods
November, 2005, Vol. 4, No. 2, 408-424

1538 – 9472/05/$95.00

A Discretized Approach to Flexibly Fit Generalized Lambda Distributions to Data

Steve Su
Epi-stat Division, George Institute for International Health
Sydney, New South Wales, Australia

This article presents a flexible approach to fit statistical distribution to data. It optimizes the bin-width of data histogram to find a suitable generalized lambda distribution. In addition to the default optimization, this approach provides additional flexibility akin to the concepts of loess and kernel smoothing, which allow the users to determine the amount of details they would like to smooth over the data. The approach presented in this article will allow users to visually compare and choose the parameters of generalized lambda distribution that best suit their purposes of study.

Key words: generalized lambda distributions, quantile distributions, fitting distributions to data

Introduction

An essential problem in data analysis is to find a probability distribution that will adequately fit the empirical data. Considerable literature exists in this area, ranging from the parametric work of generalized lambda distribution (Ramberg & Schmeriser, 1974; Ramberg, Tadikamalla, Dudewicz & Mykytka, 1979; Ozturk & Dale, 1985; Freimer, Mudholkar, Kollia, & Lin, 1988; Okur, 1988; King & MacGillivray, 1999; Karian & Dudewicz, 2000; Lakhany & Massuer, 2000) to nonparametric work of kernel density estimation (Silverman, 1985). In spite of these works, no current work exists on allowing a range of possible generalized lambda distribution (GλD) fits to data, pending on users' desire to suppress or accentuate certain features of the data based on prior knowledge of the distribution. This is important when a particular method fails to provide a fit that highlights the essential features of the data exhibited and known by the analyst. In these situations, it will often be preferable to explore other plausible GλDs.

Steve Yu Shuo Su is a Research Fellow at the Epi-stat Division of the George Institute, affiliated with the University of Sydney. His research interests are in applied statistical methods in business and epidemiology. Email: ssu@thegeorgeinstitute.org.

This article proposes an extension of the existing fitting method using GλD which offers more flexibility and in many cases can highlight features of the data not considered by the King and MacGillivray (1999)'s starship method. Instead of optimizing using goodness of fit method, this article suggests an alternative approach which is to optimize based on the number of classes or bins of the data. The number of bins of the data can be determined by the user, offering flexibility to suppress or highlight details, much like the concept of smoothing a data set using different weights in loess or kernel smoothing. This is a valuable tool in practice because the real distribution of the data set is almost never known and the methods developed in this article can be used to conduct sensitivity analysis to assess the effects of using different yet plausible distributions.

The principal emphasis in this article is to allow the user to fit a wide range of different distributions to data set rather than to satisfy the goodness of fit statistics. Also, the exclusive use of goodness of fit statistics in the fitting of distribution to data as was done in previous works (King & MacGillivray, 1999; Lakhany & Massuer, 2000) does not guarantee the resulting distribution fit will satisfy the goodness of fit, but merely tries to maximize it. The beauty of the approach in this article is that it allows the data to be represented in different angles. This is important because unlike theoretical simulated data, real life data is often messy. Very often,

real life data does not have a nice continuous range of values one can get from theoretical simulations. Due to this imperfection, it is often desirable to have an alternative data fitting method that could provide alternative fits beyond the traditional goodness of fit methods. This will give the user a possible range of distribution fits that could arise from the data set and this can lead to valuable sensitivity analysis on the impact of different distributions. The use of goodness of fit criteria could also enhance the credibility of fit under different fits but should not discredit it. This is because it is only possible to test the goodness of fit of one realization of the real life data from its underlying distribution, which may or may not be representative.

The article begins with a literature review on the existing methods of GλD parameters estimation, which progressively result in the development of this new method. Results of the application of the new methods on real life data are then presented and the article concludes with a discussion on the shortcomings of this new method.

Review of Literature

This literature review begins with the basic theory of GλD and discusses some of the fitting methods reported in literature. The literature review then presents two methods that appear to give promising results. These two methods are extended and discussed in the method section.

The Ramberg-Schmeiser (1974) (RS) GλD is an extension of Tukey's lambda distribution (Hastings, Mosteller, Tukey, & C 1947). It is defined by its inverse distribution function:

$$F^{-1}(u) = \lambda_1 + \frac{u^{\lambda_3} - (1-u)^{\lambda_4}}{\lambda_2} \qquad (1)$$

In Expression (1), $0 \le u \le 1$, $\lambda_2 \neq 0$ and λ_1, λ_2, λ_3, λ_4 are respectively the location, scale, skewness and kurtosis parameters of generalized lambda distribution GλD(λ_1, λ_2, λ_3, λ_4). In particular, Karian, Dudewicz and MacDonald (1996) noted that GλD is defined if and only if:

$$\frac{\lambda_2}{\lambda_3 u^{\lambda_3 - 1} + \lambda_4 (1-u)^{\lambda_4 - 1}} \ge 0 \qquad (2)$$
$$u \in [0,1]$$

Another distribution known as FMKL GλD also exists, due to the work of Freimer Mudholkar, Kollia and Lin (1988). This distribution is slightly different to RS GλD and they overlap when $\lambda_3=\lambda_4$. The FMKL GλD can be written as:

$$F^{-1}(u) = \lambda_1 + \frac{\frac{u^{\lambda_3} - 1}{\lambda_3} - \frac{(1-u)^{\lambda_4} - 1}{\lambda_4}}{\lambda_2} \qquad (3)$$

Under Expression (3), $0 \le u \le 1$, and λ_1, λ_2, λ_3, λ_4 are consistent with the interpretations in RS GλD, namely λ_1, λ_2 are the location and scale parameters and λ_3, λ_4 are the shape parameters. In particular, if $\lambda_3=\lambda_4=0$, both RS and FMKL GλD have:

$$F^{-1}(u) = \lambda_1 + \frac{\ln(u) - \ln(1-u)}{\lambda_2} \qquad (4)$$

The fundamental motivation for the development of FMKL GλD is that the distribution is proper over all λ_3 and λ_4 (Freimer, Mudholkar, Kollia, & Lin, 1988). This adds convenience to users who wish to program this function as there are fewer restrictions on the values of λ_3 and λ_4. The only restriction on FMKL GλD is $\lambda_2>0$.

The extensive use of FMKL GλD is reported in Freimer et al (1988). Due to the wide range of shapes GλD possesses, for example: U shaped, bell shaped, triangular, and exponentially shaped distributions and its simplicity, it has been used in Monte Carlo simulations (Hogben, 1963), the modeling of empirical distributions (Ramberg, Tadikamalla, Dudewicz, & Mykytka, 1979; Okur, 1988), and in the sensitivity analysis of robust statistical methods (Shapiro, Wilk, & Chen, 1968). Other

research works on GλD concentrate on estimating the parameters of the GλD from empirical data and these are discussed below.
In any optimization problem, it is necessary to:

1. Find suitable initial values, and
2. Choose the appropriate optimization scheme.

Perhaps the most common approach has been to use method of moments to estimate the parameters of GλD as demonstrated in Ramberg et al (1979) and Karian and Dudewicz (1996, 2000). These works covered only the RS GλD and often use tables based on the third and fourth moments or percentiles of the data to find suitable initial values. The appropriate optimization scheme involves finding a GλD with parameters λ_1, λ_2, λ_3, λ_4 that matches closely with the first four moments of the empirical data. This is done numerically through either the Nelder-Simplex (Nelder & Mead, 1965) algorithm as in the work of Ramberg, et al. (1979) or the Newton-Raphson algorithm or tabulated values (Karian & Dudewicz, 2000). Karian and Dudewicz (1996) also discussed the use of the generalized beta distribution to model the distributions that were not covered by the original RS GλD. In Karian and Dudewicz (2000), an alternative method is also demonstrated which matches the RS GλD with the parameters λ_1, λ_2, λ_3, λ_4 based on the first four percentiles of the data set. This is a variation on the same theme of the matching of moment method but one in which Karian and Dudewicz (2000) reported can produce better fits than in the case with other methods of moment matching under RS GλD.

In a different line of work, Ozturk and Dale (1985) used a version of least squares estimation to find the parameters of RS GλD. They derived the squared distance between empirical data points with the expected values of the order statistics, and numerically minimized this measure using Nelder-Simplex method to derive parameter estimates for the RS GλD.

The literature recognizes that matching the first four moments or using the "least squares" method by Ozturk and Dale (1985) does not necessarily produce a good fit to the data (Karian & Dudewicz, 2000; Lakhany & Massuer, 2000). This is due to different parameters of the GλD can results in the similar first four moments. For example, in the case of the least squares method by Ozturk and Dale (1985), the goal of minimizing the squared distance between empirical data points with the expected values of the order statistics of GλD does not necessarily coincide with the formal goodness of fit objective such as the Kolmogorov-Smirnov Goodness-of-Fit Test.

It is precisely the need to assess the resulting fit with the goodness of fit objective that King and MacGillivray (1999) used the starship methods. In the starship method, grid points comprising of λ_1, λ_2, λ_3, λ_4 aimed at covering a wide range of GλD, calculated from the sample quantiles. Then, for each of the grid points the theoretical GλD was transformed into uniform distribution and goodness of fit statistics like Anderson-Darling test statistics or Kolmogorov-Smirnov test statistics were calculated. The set of grid points with the lowest Anderson-Darling statistics was then being chosen as the initial values for optimization, usually through the Nelder-Simplex algorithm. The resulting values from the optimization scheme are the parameter estimates of the GλD, given by starship method.

Lakhany and Mausser (2000) suggested a variation of using re-sampling method combined with the method of moments and a goodness of fit test via the FMKL GλD. They first generated initial values for the method of moment matching via quasi random number generator (i.e., the Sobol sequence generator (Bratley & Fox, 1988)), and then found the set of values λ_1, λ_2, λ_3, λ_4 that matched optimally (through the Nelder-Simplex algorithm) with the first four moments from the data. This set of values was then evaluated through a goodness of test statistic such as adjusted Kolmogorov-Smirnov test statistics. Under this method, any solution that results in a p-value > 0.05 is accepted. Lakhany and Mausser (2000) commented that this method is much more efficient time-wise than the starship method developed by King and MacGillivray (1999) and allows for automatic restarts from different initial values to help to find a distribution that will adequately fit the data. The use of p-values in the optimization scheme, however, can be

somewhat problematic. The deficiency of p-values is well known, since failure to reject does not mean the hypothesis is true since it may be that the sample size is too small to be able to detect differences between the empirical and fitted data. Conversely, rejection of the hypothesis does not mean the fitted model is inappropriate, as the user may have a different purpose to fitting the data other than to satisfy the goodness of fit criteria.

An important improvement of Lakhany and Mausser (2000)'s approach is the flexibility of fits it offers to the users. As different initial values are chosen, different results can be obtained. However, this flexibility is rather limited as the users have no real control over the amount of smoothing they would like to achieve.

The current literature does not appear to cover a comparison of the method of percentiles from Karian and Dudewicz (2000) with the other methods like starship by King and MacGillivray (1999), nor with the automatic re-sampling methods of Lakhany and Massuer (2000). The method below will consider both the method of percentiles under RS GλD and the method of moments under the FMKL GλD. The rationale is that the existing literature appears to recommend these two methods hence these methods are chosen for extension to offer greater flexibility of fit than the methods previously reported.

A detailed discussion of the method of percentiles using the RS GλD and the method of moments using FMKL GλD is outlined below.

Method of percentiles using the RS GλD:

The following is obtained directly from Karian and Dudewicz (2000). For a given data set X with values x_1, x_2, x_n, the p-th percentile defined by Karian and Dudewicz (2000) is $\hat{\pi}_p = y_r + k(y_{r+1} + y_r)$, where Y= y_1, y_2,… y_n are sorted values of X in ascending order and r is the truncated value of (n+1)×p with k being (n+1)×p-r.

Instead of using the first four moments, the following statistics are used:

$$\begin{aligned}
\hat{\rho}_1 &= \hat{\pi}_{0.5} \\
\hat{\rho}_2 &= \hat{\pi}_{1-v} - \hat{\pi}_v \\
\hat{\rho}_3 &= \frac{\hat{\pi}_{0.5} - \hat{\pi}_v}{\hat{\pi}_{1-v} - \hat{\pi}_{0.5}} \\
\hat{\rho}_4 &= \frac{\hat{\pi}_{0.75} - \hat{\pi}_{0.25}}{\hat{\rho}_2}
\end{aligned} \tag{5}$$

where v is an arbitrary number from 0 to 0.25.

The relationship between the theoretical ρ_1, ρ_2, ρ_3, ρ_4 and λ_1, λ_2, λ_3, λ_4 in the RS GλD is as follows:

$$\begin{aligned}
\rho_1 &= F^{-1}(0.5) = \lambda_1 + \frac{0.5^{\lambda_3} - 0.5^{\lambda_4}}{\lambda_2} \\
\rho_2 &= F^{-1}(1-v) - F^{-1}(v) = \frac{(1-v)^{\lambda_3} - v^{\lambda_4} + (1-v)^{\lambda_4} - v^{\lambda_3}}{\lambda_2} \\
\rho_3 &= \frac{F^{-1}(0.5) - F^{-1}(v)}{F^{-1}(1-v) - F^{-1}(0.5)} = \frac{(1-v)^{\lambda_4} - v^{\lambda_3} + (0.5)^{\lambda_3} - (0.5)^{\lambda_4}}{(1-v)^{\lambda_3} - v^{\lambda_4} + (0.5)^{\lambda_4} - (0.5)^{\lambda_3}} \\
\rho_4 &= \frac{F^{-1}(0.75) - F^{-1}(0.5)}{\rho_2} = \frac{(0.75)^{\lambda_3} - (0.25)^{\lambda_4} + (0.75)^{\lambda_4} - (0.25)^{\lambda_3}}{\rho_2}
\end{aligned} \tag{6}$$

The condition $-\infty < \rho_1 < \infty, \rho_2 \geq 0, \rho_3 \geq 0, \rho_4 \in [0,1]$ must also be true, which is a direct consequence of the definition of ρ_1, ρ_2, ρ_3, ρ_4. In Karian and Dudewicz (2000), a fit for the GλD is found by solving Expression (7) through the use of tables. This can also be solved this numerically via Newton-Raphson method.

$$\left|\hat{\rho}_3 - \rho_3\right| <= 10^{-6}, \left|\hat{\rho}_4 - \rho_4\right| <= 10^{-6} \tag{7}$$

In the extended method described below, however, the following minimization scheme in Expression (8) is used. Once λ_3, λ_4 are obtained, λ_1, λ_2 can be obtained directly via Expression (6).

$$\sqrt{\left(\hat{\rho}_3-\rho_3\right)^2+\left(\hat{\rho}_4-\rho_4\right)^2} \tag{8}$$

Method of Moments under the FMKL GλD:

In an alternative approach, Lakhany and Mausser (2000) used the method of moments for the FMKL GλD. The following are extracts from Lakhany and Mausser (2000):

For a given data set X with values x_1, x_2,… x_n, the i-th moment α_i is defined in Expression (9).

$$\begin{aligned}
\hat{\alpha}_1 &= \frac{\sum_{i=1}^{n} x_i}{n} \\
\hat{\alpha}_2 &= \frac{\sum_{i=1}^{n} (x_i-\hat{\alpha}_1)^2}{n} \\
\hat{\alpha}_3 &= \frac{\sum_{i=1}^{n} (x_i-\hat{\alpha}_1)^3}{n(\hat{\alpha}_2)^{1.5}} \\
\hat{\alpha}_4 &= \frac{\sum_{i=1}^{n} (x_i-\hat{\alpha}_1)^4}{n(\hat{\alpha}_2)^{2}}
\end{aligned} \tag{9}$$

Putting a= $\frac{1}{\lambda_2}$ and b= $\lambda_1-\frac{1}{\lambda_1\lambda_2}+\frac{1}{\lambda_2\lambda_4}$, with Y=(X-b)/a, using $E(X^k)=\int_0^1 \left(F^{-1}(u)\right)^k du$ and binomial expansion gives Expression (10).

$$\begin{aligned}
s_k &= E(Y^k) \\
s_k &= \int_0^1 \left(\frac{u^{\lambda_3}}{\lambda_3}-\frac{(1-u)^{\lambda_4}}{\lambda_4}\right)du \\
s_k &= \int_0^1 \sum_{j=o}^{k}\binom{k}{j}(-1)^j\left(\frac{u^{\lambda_3(k-j)}}{\lambda_3^{\,k-j}}-\frac{(1-u)^{\lambda_4 j}}{\lambda_4^{\,j}}\right)du \\
s_k &= \sum_{i=o}^{k}\binom{k}{j}\frac{(-1)^j}{\lambda_3^{\,k-j}\lambda_4^{\,j}}\beta(\lambda_3(k-j)+1,\lambda_4 j+1)
\end{aligned} \tag{10}$$

In Expression (10), β(*) denotes beta function. Note that both arguments of the beta function must be positive, implying that min(λ_3, λ_4) > -1/*k* if the distribution is to have finite k-th moments. The k-th central moment (except for the first which is the mean) of the distribution $F^{-1}(u)$ denoted as μ_k are hence given in Expression (11).

$$\begin{aligned}
\mu_1 &= \frac{1}{\lambda_2}(s_1)-\frac{1}{\lambda_2\lambda_3}+\frac{1}{\lambda_2\lambda_4} \\
\mu_2 &= \frac{1}{\lambda_2}(s_2-s_1^2) \\
\mu_3 &= \frac{1}{\lambda_2^3}(s_3-3s_1s_2+2s_1^3) \\
\mu_4 &= \frac{1}{\lambda_2^4}(s_4-4s_1s_3+6s_1^2s_2-3s_1^4)
\end{aligned} \tag{11}$$

The theoretical α_3 and α_4 are given in Expression (12).

$$\begin{aligned}
\alpha_3 &= \frac{s_3-3s_1s_2+2s_1^3}{(s_2-s_1)^{\frac{3}{2}}} \\
\alpha_4 &= \frac{s_4-4s_1s_3+6s_1^2s_2-3s_1^4}{(s_2-s_1)^2}
\end{aligned} \tag{12}$$

The same methodology now follows as from Lakhany and Mausser (2000). They propose to find λ_3, λ_4 by minimizing Expression (13), where $\hat{\alpha}_3$ and $\hat{\alpha}_4$ are sample values using sample moments.

$$\sqrt{\left(\hat{\alpha}_3-\alpha_3\right)^2+\left(\hat{\alpha}_4-\alpha_4\right)^2} \tag{13}$$

Once λ_3, λ_4 is determined it is possible to find λ_1, λ_2 as shown in Expression (14).

$$\lambda_2=\frac{\sqrt{\left(s_2-s_1^2\right)}}{\hat{\alpha}_2}$$
$$\lambda_1=\hat{\alpha}_1+\frac{1}{\lambda_2}\left(\frac{1}{\lambda_3+1}-\frac{1}{\lambda_4+1}\right) \tag{14}$$

Extension of previous methodology

The principle underlying earlier methods (King & MacGillivray, 1999; Lakhany & Massuer, 2000) is to use goodness of fit as a criteria to determine whether the resulting GλD fits the data adequately. However this, as will be demonstrated later, does not give the potential for a wide range of different plausible distribution fits to data.

The new method described here uses the percentile method from Karian and Dudewicz (2000) and the method of moments with the FMKL GλD. It also uses quasi random numbers to find initial values, but the optimization can be based on the number of classes or bins the user specifies. This optimization scheme allows users to suppress or accentuate part of the distribution as desired, a feature that is not explicitly considered in other methods. The range of initial values should be chosen based on the shape of the distribution shown by the histogram, or they maybe left unspecified with a default set of values chosen.

A full description of the algorithm is provided below:

1. Specify a range of initial values for λ_3, λ_4, and the number of initial values to be selected. Here, the λ_3, λ_4 are set by default to range from -1.5 to 1.5 for the RS GλD percentile method and -0.25 to 1.5 for the FMKL GλD method of moment. These default values are from author's clinical experiences and appear to work well in most situations. It is possible to change these initial values if desired.

The quasi random generator used is based on the work of Hong and Hickernell (http://www.mcqmc.org/Software.html) and the scrambling method of Owen (1995) and Faure and Tezuka (2000). This code is available from the beta resample library in Splus 6.0 and scrambling methods are applied so that the numbers generated fills uniformly onto the λ_3, λ_4 two dimensional space. To increase the speed, it is possible to set the initial values where λ_3= λ_4. This appears to work well in many situations. By default, 100 of such initial values are chosen in this case and used in step 2.

2. Evaluate λ_1, λ_2 for each of the initial values λ_3, λ_4. Remove all the set of values that do not:
    a. Result in a legal parameterization of GλD.
    b. Span the entire region of the data set.

From these sets of initial points, find the values of λ_3, λ_4 that matches closely with the data. This is to generate a set of initial values that produce the lowest values in Expression (8) and Expression (13), to be used as initial values in the optimization process.

3. Sort the sample data in ascending order, and divide the data set into evenly spaced classes with bin edges that span

the data set. Calculate the proportion of the sample out of the total sample in each class. Hence Table 1 maybe constructed:

Table 1 Calculating proportion of data in each class

Classes	1.5-2	2-2.5	2.5-3	3-3.5	Sum
Proportion of data	0.1	0.6	0.2	0.1	1

Table 1 shows four classes, with the proportion of the data set falling in each class in the second column. Let the proportion of data in each class be denoted d_i for i=1,2,3..n classes and the proportion of data from the theoretical GλD be the vector t_i for i=1,2,3…n classes. The quantity that one wants to minimize is:

$$\sum_{i=1}^{n} d_i (d_i - t_i)^2 \qquad (15)$$

Expression (15) is the weighted squared deviation of the theoretical proportions with the actual data proportions. This is weighted so that the data with higher proportions are given priority in the minimization scheme. The resulting fit will then be more likely to capture the majority of the data. The weighting factor d_i can be removed if desired. In addition, this optimization scheme also rejects estimations that do not span the entire data set.

The number of classes, n, can be solely determined by the user, or determined by the formula devised by this article (discussed below), or via previous literature works as in Sturges, Scott (1979; 1992) or Freedman and Diaconis (1981).

Sturges' formula is based a bin width of:

$$\text{range(data)}/(\log_2 \text{m} + 1) \qquad (16)$$

This strategy often results the bin width being too wide as reported in Venables and Ripley (2002), and has the disadvantage that "outliers may inflate the range and increase the bin width in the centre of the distribution."

Hyndman (1995) also argued that the use of Sturges' formula should be avoided since there is no sound statistical backing to its derivation.

Scott (1979) used $3.5\hat{\sigma} m^{-1/3}$, although Freedman & Diaconis (1981) proposed $2Rm^{-1/3}$, where R is the inter-quartile range and $\hat{\sigma}$ is the estimated standard deviation from the data, and m is the number of observations in the data. Freedman & Diaconis's (1981) use of inter-quartile range is more robust against outliers and tends to choose smaller bins than the formula by Scott (1979). More complicated rules are also available in Scott (1992) but they are not discussed here.

The methods developed in this article calculate the default number of classes to be optimized over as the one that gives ζ: the minimal squared error between the first two moments of the categorized data with the actual. For example, in the context of Table 1, the first two moments of the categorized data can be calculated using the following table, which takes the mid point of the class intervals and treat the data as discrete. The mean and variance of data shown in Table 2 are 2.4 and 0.1525 respectively; this is then compared with the actual mean and variance of the continuous data with the squared error subsequently calculated. The number of classes chosen for optimization would be the one with minimal squared error or ζ. It is possible to choose any other number of classes such as the formula in Scott (1979) and Freedman & Diaconis (1981).

Table 2 Calculating mean and variance from Table 1

Observation	1.75	2.25	2.75	3.25	Sum
Proportion of data	0.1	0.6	0.2	0.1	1

The philosophy for this approach is to choose the number of classes that best represents the first two moments of the data, so that the distribution fitted would resemble more or less an accurate representation of the data set.

Although formulas for determining the optimal bin width for the histograms interval do exist, users can exercise their judgments by choosing the number of classes. Generally

speaking, higher number of classes will result in details of the distribution being accentuated, while lower number of classes will tend to suppress details of the distribution.

4. The optimal result can be obtained via the Nelder-Mead Simplex algorithm or another suitable numerical optimization algorithm. It is advisable to re-use the initial values in the optimization process to ensure the result obtained is a global minimum rather than a local minimum. Steps 1 to 3 may be repeated if necessary, where the number of classes and the range of initial values can be adjusted until the results are deemed adequate. The final fitting result can be examined by plotting the result on the histogram with the fitted line as well as testing the goodness of fit using the Kolmogorov-Smirnov (KS) test.

Results

The analysis below is divided into two parts. The first part is a theoretical comparison between data fitting methods with well known statistical distributions. A two sample KS test is carried out by sampling 100 points from the theoretical and fitted distributions and the number of times the p-value exceeds 0.05 is recorded over 1000 times. This will give the user an independent measure as to the adequacy of fits beyond a visual comparison. The second part shows the fitting method over some real life data, and the goodness of fit test is carried out on the comparison between sampling 90% of the real life data with the fitted data using two sample KS test over 1000 runs.

This is also known as the Monte Carlo KS test in this article. It is worth cautioning that the use of goodness of fit as a measure for quality of fit would bias methods that seek to maximize goodness of fit. In fact, it is a circular logic. The use of goodness of fit to assess the quality of fits used in this article will not suffer from this problem, but it needs to bear in mind that the objective of fit in this article was not to maximize the goodness of fit, and so it may not always be as high as starship method (STAR) which uses standard statistical goodness of fit such as Kolmogorov-Smirnov and Anderson Darling test statistics in its data fitting algorithm.

The following compares between the revised percentile method of the RS GλD (RPRS), the revised method of moment under the FMKL GλD (RMFMKL) and the STAR method. Previous literature such as King and MacGillivray (1999), Lakhany and Mausser (2000), and Karian and Dudewicz (2000) have already covered comparisons between the starship methods, the GλD under the RS and FMKL GλD using the method of moments and percentiles as well as the least square method used by Ozturk (1985); hence these will not be repeated here.

Commentary

The modified methods RPRS and RMFMKL are perhaps not appropriately termed as the percentiles and method of moments are not used in the optimization step but only for choosing the initial values for the optimization process. However, the differences in the two methods highlight the fact that the choices of initial values and type of GλD are important in the outcome of these extended methods, since different results are obtained even though both methods undergo the same optimization scheme.

Comparison with Theoretical Distributions

Figure 1 and Table 3 show the resulting fits of RPRS, RMFMKL and STAR on well known statistical distributions. Using the default fitting method described above, RPRS and RMFMKL are very close to the actual distribution in Figure 1. This result is further confirmed in Table 3, where more than 90% of the time, the Monte Carlo KS test will indicate there is no difference between the fitted and actual distributions.

The real interest of the method of this article is not in the fitting of theoretical distributions. In the theoretical simulation it is possible to compare between the actual and approximate distributions, but not so in practice. It is precisely the reason that one does not know the real underlying distribution of real life data, one needs a flexible fitting method that could allow us to assess different distribution fits and

the stability of distribution fits under different data representations by the histogram.

The following real life examples will compare different cases where different methods work well under different situations. It will also use the Monte Carlo KS tests results to demonstrate the quality of fit under the goodness of fit objective.

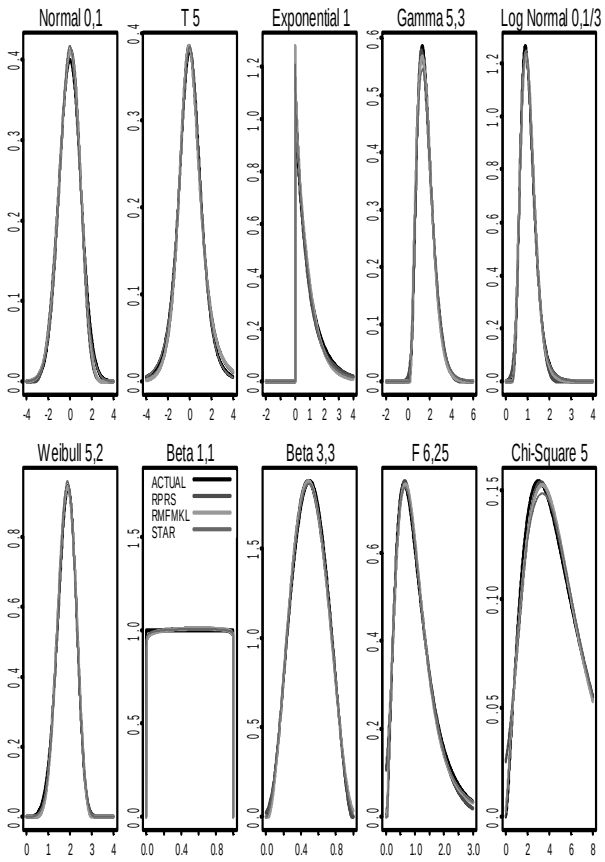


Figure 1: Demonstrating the distribution fits of well known statistical distributions.

Table 3: Monte Carlo KS goodness of fit tests results over 1000 runs. A value close to 1000 indicates high level of confidence of a good fit.

Distribution	RPRS	RMFMKL	STAR
normal(0,1)	941	966	955
student(5)	943	940	960
exp(1)	945	905	944
gamma(5,3)	957	960	961
lognormal(0,	967	977	969
weibull(5,2)	964	968	952
beta(1,1)	970	963	970
beta(3,3)	966	966	959
f(6,25)	939	964	961
chisq(5)	962	966	958

Dataset used

The datasets used in here were supplied by research works of Sabri Hassan and Victoria Clout at School of Accountancy in Queensland University of Technology, Australia. The dataset by Sabri Hassan is based on 44 Australian extractive industries firms, listed on the ASX (Australian Stock Exchange) from 1998 to 2001. The dataset used is based on the mean value of each individual company over four years. Market to Book values (sh.mtb), transparency (sh.transp), and profit (sh.profit) variables were extracted and used in this demonstration. There are 176 observations in this data set and the goodness of fit test below will sample 160 observations from this data set and the fitted distribution.

Victoria Clout's data consisted of 361 US firms, listed on the S&P500. The selection requirements were December year-end firms for the 1977 to 1995 period. Similarly, the data used is based on the mean values for each company over the 12 years period. Market to Book ratio (vc.mbr), Ratio of cash and marketable securities over current assets (vc.flex), return on assets (vc.roa) were used in this demonstration. There are 143 observations in this data set and the goodness of fit test below will sample 130 observations from this data set and the fitted distribution.

In addition to financial data, geological data (faithful) on the duration of 272 eruptions

from the Old Faithful geyser in Yellowstone National Park (Hardle, 1991) was also used.

The following examples are designed to demonstrate the flexibility the new methods which can fit alternative, convincing distributions other than suggested by the starship method. It also designed to offer a balanced view on some of the possible deficiencies of this method in relation to satisfying the goodness of fit tests.

Figure 2 is an example of graphical over-fitting by the STAR method, and how the use of default settings described in this article appears to give a more adequate fit. The number of classes to be optimized over is 12, using the default calculations. The histogram shown in Figure 2 is plotted using 100 classes. Using the Monte Carlo KS test, the results are 0, 7 and 732 for RPRS, RMFMKL and STAR respectively. This suggests that STAR is the best fit among the three under the Monte Carlo KS test. It is however possible to improve the Monte Carlo KS test of the RPRS fit by increasing the number of classes to be fitted.

Example 1: sh.mtb

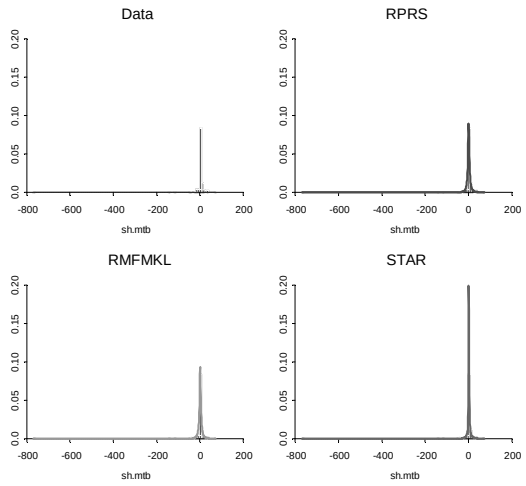


Figure 2: Fitting of sh.mtb data using RPRS, RMFMKL and STAR methods. The extreme scale is due to an extreme outlier, which is retained for illustrative purposes. For example, a certain process may have a huge loss with a very small probability, but it is nevertheless important to model that scenario.

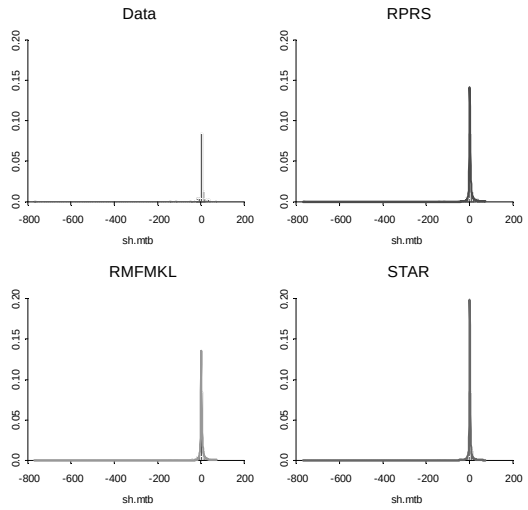
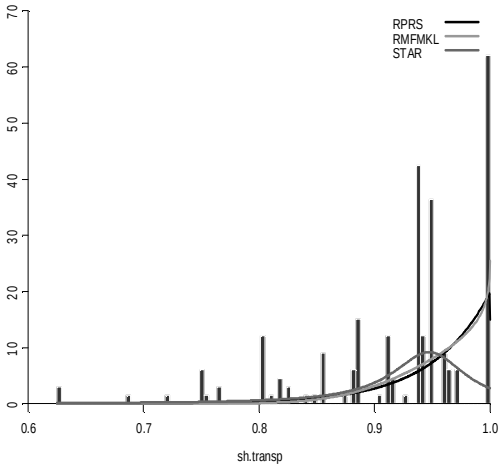


Figure 3: Fitting of sh.mtb data using RPRS, RMFMKL and STAR methods using 150 classes. This shows how it is possible to fit using different histogram bin widths to improve the goodness of fit.

Figure 3 shows the result of such fit graphically and the Monte Carlo KS results are 585, 561 and 749 for RPRS, RMFMKL and STAR. A real strength of the method developed in this article is that it gives a range of plausible fits which the goodness of fit could be assessed objectively. For example, it can be considered that the results in Figure 2 are less likely to be the real representation of the data than Figure 3.

Example 2: sh.transp, alternatives suggested by RPRS, RMFMKL:



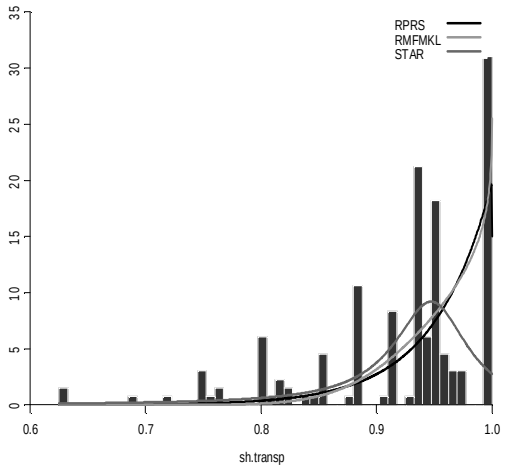


Figure 4: Figures showing fitting of sh.transp data using RPRS, RMFMKL and STAR, the first histogram uses 100 classes while the second histogram uses 50 classes.

The graphs in Figure 4 show two histograms with 100 and 50 classes with the default optimization classes to be optimized over being 31. STAR failed to capture the upward trend of the data. If it is desirable to reach the peak of the histogram data with 100 classes, it is possible to refit RPRS and RMFMKL over 100 classes, resulting in Figure 5. Using 50 or 100 classes will result in Monte Carlo KS test results of 0, 0, and 300 for RPRS, RMFMKL and STAR.

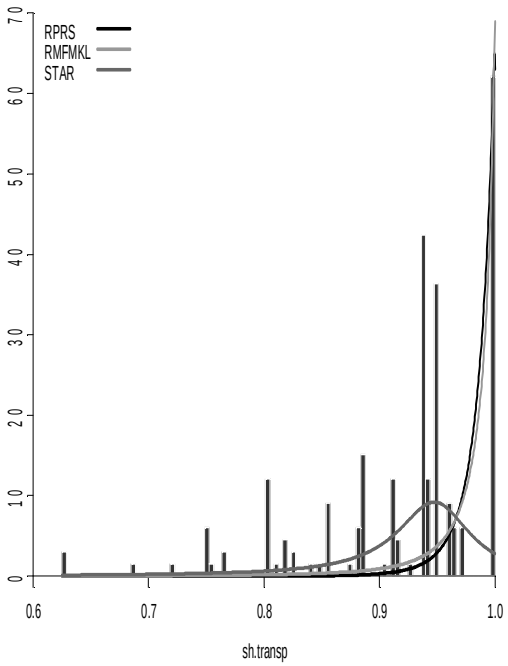


Figure 5: Figure showing alternative fitting of sh.transp sh.transp by RPRS and RMFMKL using 100 histogram classes.

This suggests that none of the methods appear to work well in this case, as STAR although the best out of the three in the Monte Carlo KS test, only really can be said to represent the data 3 times out of 10. In situation like this, where none of the method appears to work well, it is useful to explore other plausible fits and conduct sensitivity analysis to examine the impact on a particular analysis using different distributions.

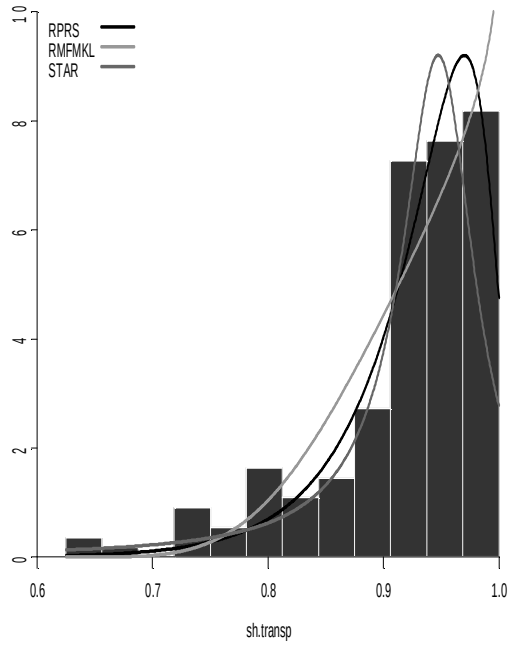


Figure 6: Figure showing alternative fitting of sh.transp using 12 histogram classes.

Figure 6 shows how STAR captured a different representation of the dataset; by manually adjusting the classes of histograms to 12, the fit by STAR appears to be more plausible. Alternative fits by RPRS and RMFMKL using 12 classes appears to represent the data well. This example highlights the importance of allowing alternative methods, since they can give different and possibly valid representations to the same data set. The Monte Carlo KS test results are 23, 2 and 290 for RPRS, RMFMKL and STAR. It also shows the flexibility of RPRS and RMFMKL which can give different fits to the data set depending on the number of classes specified. An additional analysis showing the effect of changing number of classes from 5 to 55 and the corresponding RPRS and RMFMKL fits is shown in Figure 7. All the Monte Carlo KS test results under each of the class suggest 0, 0 and 300 for RPRS, RMFMKL and STAR respectively. The graphs show how different fits may be obtained by varying the number of classes and it is possible these may not change the result of the Monte Carlo KS tests at all. The sharp spike exhibited in Figure 7 for 15 classes is characteristic of RPRS fits, as will be shown in more examples below.

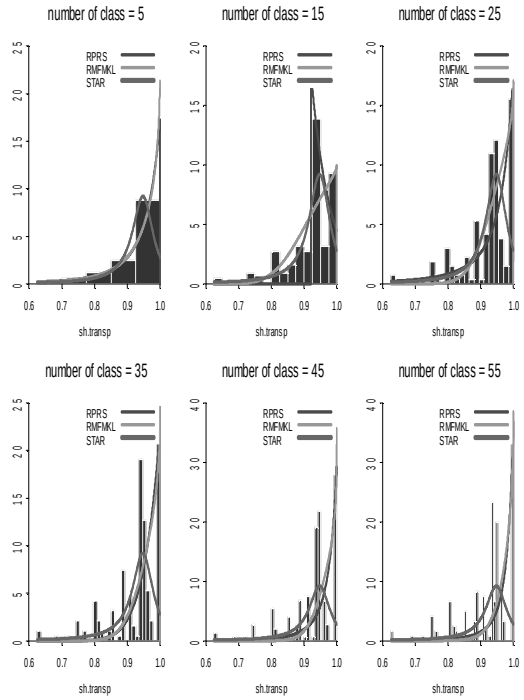


Figure 7: Figure showing alternative fitting of sh.transp using different histogram classes.

Example 3: vc.leverage, similar results:

This example shows that consistent results can often be obtained between different methods. RPRS and RMFMKL used 89 classes by default calculations in this case. The result is shown in Figure 8 below with the histogram exhibiting 100 classes. The Monte Carlo KS tests suggest 882,887 and 945 for RPRS, RMFMKL and STAR respectively. It is normally the case that STAR has somewhat higher goodness of fit score, owing to its fitting objective.

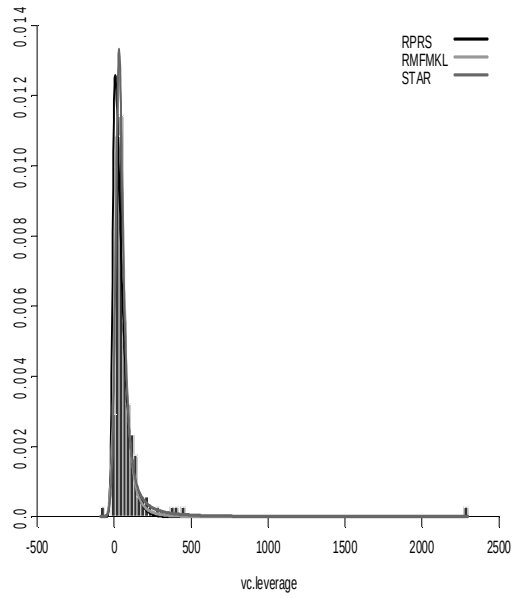


Figure 8: Figure showing fitting of vc.roa data using RPRS, RMFMKL and STAR. All methods give similar results.

Example 4: vc.mbr

RPRS and RMFMKL used 20 classes by default calculations in this optimization scheme. Figure 9 shows a histogram with 100 classes, and all methods give different representations to the dataset. They are all valid representations as suggested by Monte Carlo KS tests, with 929, 887 and 934 for RPRS, RMFMKL and STAR. A striking feature is that RPRS is similar to RMFMKL and they appear to capture the peak of data better than the STAR method. An additional analysis showing the effect of changing number of classes from 5 to 55 and the corresponding RPRS and RMFMKL fits is shown in Figure 10. This example shows how plausible fits can be gauged by using the method described in this article. Table 4 shows the resulting Monte Carlo KS tests for different number of classes and it can be used to as a rough guide as to how credible certain fits are to the data set. For example, for RMFMKL, the most plausible fits are from classes of 15 and 35. This example at Table 4 also shows that the method developed in this article can be as good as STAR method, in addition to offering flexibility to provide convincing fits.

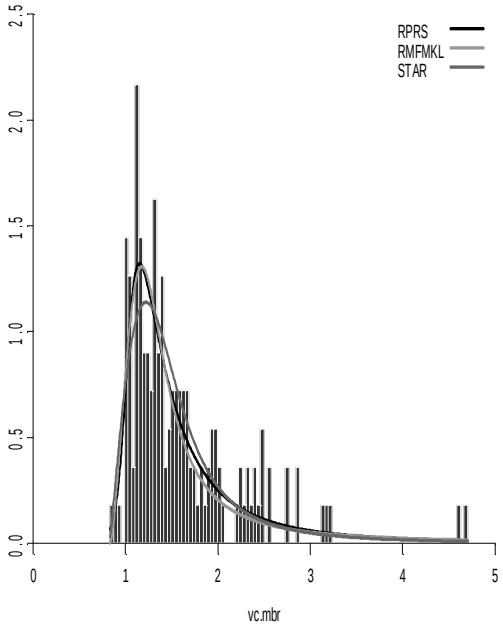


Figure 9: Figure showing fitting of vc.mbr data using RPRS, RMFMKL and STAR. RPRS and RMFMKL appear to represent the peak of the data better than STAR.

Table 4: Monte Carlo KS test for vc.mbr over different number of classes

Classes Method	5	15	25	35	45	55
RPRS	481	940	933	905	908	873
RMFMKL	354	929	713	932	812	778
STAR	932	930	923	917	942	925

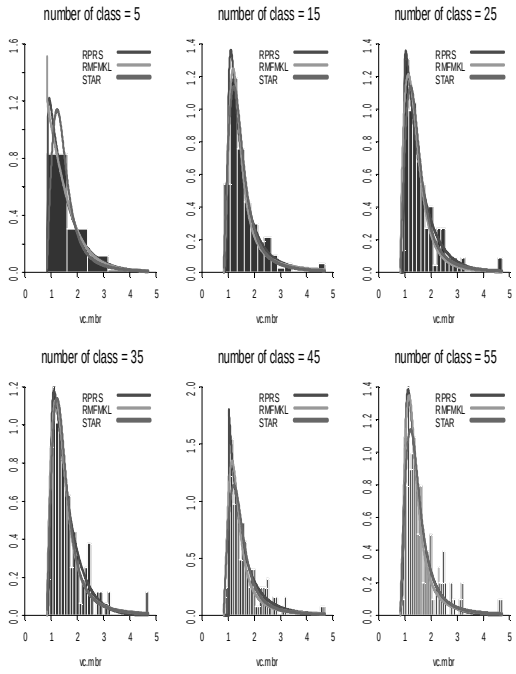


Figure 10: Figure showing alternative fitting of vc.mbr using different histogram classes.

Example 5: faithful, bimodal data, splitting fits by STAR, RPS and RMFMKL:

This last example shows cases where it may be difficult to fit the data adequately when one encounters a bimodal shaped data. In such cases, the data can be divided into two with two different distributions fitted on each side. Problem can arise when the end points do not match as appeared to be possible with the STAR method in this case. However, as shown in Figure 11, this can be easily corrected for example, by setting the optimization scheme to only include distributions that have maximum values less or equal to 3 for the distribution on the left hand side, and the distribution to have minimum values bigger or equal to 3 on the right hand side.

The original default number of classes was 52 on the RHS of Figure 11 and it does not satisfy the Monte Carlo KS test well, with 614 and 187 for RPRS and RMFMKL. Instead of using the default class calculation, the number of classes was manually adjusted to 20 and this result in Monte Carlo KS test of 855, 873 and 890 for RPRS, RMFMKL and STAR. On the LHS the default setting of 15 classes satisfy the Monte Carlo KS test well, resulting in 921, 927 and 917 for RPRS, RMFMKL and STAR and very similar fits. Figure 11 shows three plausible alternative fits and it is possible some data set may require a mixture of RS and FMKL GλD. The alternative fit by KDE is also provided in Figure 12 for comparison purposes. Figure 12 shows two different fits using KDE. However, the KDE fit, in an attempt to reach the more extreme points of the histogram became less smooth. This rugged appearance will not occur from using generalized lambda distributions.

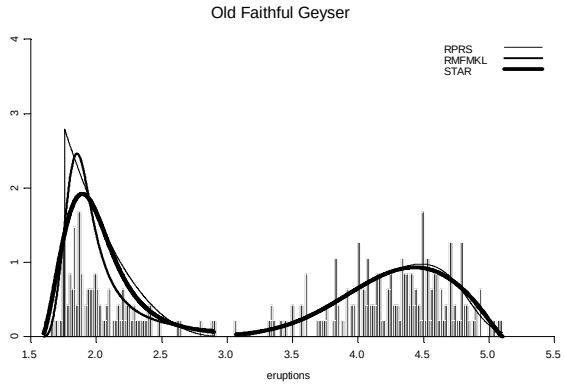


Figure 11: Figure showing fitting of eruptions data using RPRS, RMFMKL and STAR and the use of splitting techniques in fitting bi-modal shaped data. The values below 3 are fitted first and the values above 3 are fitted later.

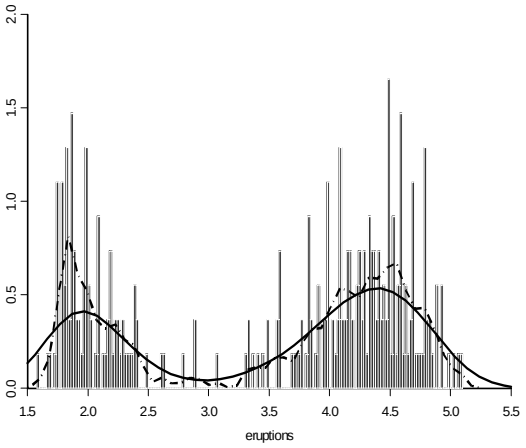


Figure 12: Graph showing two different KDE fits for the eruptions data.

Application of fitting distributions to data using GλD, and a comparison to Kernel Density Estimation method

The use of RPRS or RMFMKL can help users to model a wide variety of distributions as well as acting as a smoothing device with the flexibility of increasing or decreasing levels of details of the data. Another method that allows for density estimation is Kernel Density Estimation (KDE) (Silverman 1985). This is a nonparametric method of estimating the distribution of the data and can often result in a rather rugged appearance compared to the smooth fits from using GλD. Another advantage of using GλD is that the parametric form of the function is known. Consequently, mathematical analysis on the function is possible. In considering re-sampling from the modeled distributions for simulation purposes, both KDE and GλD could be used.

Simulation from KDE and GλD

Simulation from KDE is a simple exercise. KDE calculations give k sets of $(x_1,y_1)\ldots (x_k,y_k)$ co-ordinates which span the distribution of the data. For each consecutive set of points, the area under the line is a trapezium. Let this area be $t_1, t_2,\ldots t_{k-1}$.

Assume one want to sample n numbers from the KDE distribution. For each of the interval i=1,2,3,… k-1, calculate nt_i, and generate nt_i numbers from a uniform distribution on the interval, repeating the process for all k-1 intervals.

Simulation from GλD simply requires generating n uniform distribution over [0, 1] and substituting the result into Expression (1) for the RS GλD and Expressions (3) for the FMKL GλD.

Shortcomings of the RPRS AND RMFMKL

All methodologies have their shortcomings, and the method devised here is no exception. The design of the RPRS and RMFMKL can suffer from the following deficiencies.

1. Different results in different runs for the same settings. RPRS and RMFMKL is based on re-sampling methods over the specified range of initial values, hence different runs will result in different initial values being chosen. This is the reason sampling is based on scrambled quasi random sampling (Owen 1995; Hong & Hickernell, 2002) available from the Splus beta resample library, so that the values span evenly throughout the ranges each time. In most cases there are no dramatic changes between each run; however situations do occur when the one run results in a better fit than other runs. This problem can be minimized by increasing the number of values to be sampled in the region. For example, if one million points were chosen over the span of [-1.5, 1.5] then dramatic changes in the result between different runs would be less likely.

2. Optimization method converges falsely or do not converge. This is a problem associated with all numerical optimization schemes, rather than related to this method directly. The program written for RPRS and RMFMKL allows for the quasi-Newton method, conjugate gradients method (Fletcher & Reeves, 1964), the Nelder-Mead algorithm (Nelder & Mead, 1965) and SANN (Belisle, 1992). Hence if one optimization method fails, the other methods can be used instead. So far the use of Nelder-Mead algorithm has proven to be effective in the cases examined here and no case of non convergence have occurred in the application of this optimization procedure.

3. Subjective choice of the number of classes required. Considerable difficulties can arise when choosing number of classes for optimization. While this flexibility is intended, it also may allow data analysts to manipulate the results and choose a method that appears to suit their needs, rather than one that is the most representative of the data. This deficiency does not affect the starship method, which only allows one optimal output based on the goodness of fit measure.

Conclusion

The exposition in the result section shows the methods developed in this article can offer good alternatives of fitting distribution to data in terms of satisfying Monte Carlo KS tests. While the use of RPRS and RMFMKL offers great flexibility, it also offers rooms for subjective bias in selecting the adequate fit. The use of goodness of fit statistics, however, can help the user to determine the likelihood of a certain distribution fit in the absence of expert knowledge of the underlying data set.

In some situations, where the goodness of fit statistics cannot be adequately satisfied the user could use the methods developed in this article to conduct sensitivity analysis on the impact of results using different distributions. Lastly, improvement on the current RPRS and RMFMKL is also possible by at least two ways, by either improving the optimization algorithm or set an algorithm to quickly find plausible initial values.

References

Belisle, C. J. P. (1992). Convergence theorems for a class of simulated annealing algorithms on R^d. *Journal of Applied Probability, 29*, 885-895.

Bratley, P., & Fox, B. (1988). Algorithm 659, Implementing Sobol's Quasirandom Sequence Generator. *ACM Transactions on Mathematical Software, 14*(1), 88-100.

Faure, H., & Tezuka, S. (2000). Another random scrambling of digital (t,s)-sequences. *MCQMC 200*0, Hong Kong: Springer-Verlag.

Fletcher, R., & Reeves, C. M. (1964). Function minimization by conjugate gradients. *Computer Journal, 7*, 148-154.

Freedman, D., & Diaconis, P. (1981). On the histogram as a density estimator: L_2 theory. *Zeitschrift für Ahrscheinlichkeitstheorie und Verwandte Gebiete, 57*, 453–476.

Freimer, M., Mudholkar, G., Kollia, G., & Lin, C. (1988). A Study of the generalised Tukey lambda family. *Communications in Statistics- Theory and Methods, 17*, 3547-3567.

Hardle, W. (1991). *Smoothing Techniques with Implementation in S*. New York: Springer.

Hastings, J. C., Mosteller, F., Tukey, J. W., & C, W. (1947). Low moments for small samples: A comparative study of order statstics. *The Annals of Statistics, 18*, 413-426.

Hogben, D. (1963). *Some Properties of Tukey's Test for Non-Additivity*. NJ: Rutgers, The State University of New Jersey.

Hong, H. S. & Hickernell, F. J. (2002). *Implementing scrambled digital sequences*. Unpublished.

Karian, Z., & Dudewicz, E. (2000). *Fitting statistical distributions: The generalized lambda distribution and generalized bootstrap methods*. New York: Chapman & Hall.

Karian, Z., Dudewicz, E., & McDonald, P. (1996). The extended generalized lambda distribution systems for fitting distributions to data: History, completion of theory, tables, applications, the "final word" on moment fits. *Communications in Statistic: Computation and Simulation, 25*(3), 611-642.

King, R., & MacGillivray, H. (1999). A starship estimation method for the generalised lambda distributions. *Australia and New Zealand Journal of Statistics, 41*(3), 353-374.

Lakhany, A., Massuer, H. (2000). Estimating the parameters of the generalised lambda distribution. *Algo Research Quarterly, 3*(3), 47-58.

Nelder, J. A., & Mead, R. (1965). A simplex algorithm for function minimization. *Computer Journal, 7*, 308-313.

Okur, M. (1988). On fitting the generalised lambda distribution to air pollution data. *Atmospheric Environment, 22*, 2569-2572.

Owen, A. (1995). Randomly permuted (t,m,s)-nets and (t,s)-sequences. In (H. Niederreiter& P. J. Shiue, Eds.) *Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*, NY: Springer-Verlag. *106***,** 299-317.

Ozturk, A., & Dale, R. (1985). Least squares estimation of the parameters of the generalised lambda distribution. *Technometrics, 27*, 8-84.

Ramberg, J., & Schmeriser, B. (1974). An approximate method for generating asymmetric random variables. *Communications of the Association for Computing Machinery, 17*, 78-82.

Ramberg, J., Tadikamalla, P., Dudewicz, E., Mykytka, E. (1979). A probability distribution and its uses in fitting the data. *Technometrics, 21*, 201-214.

Scott, D. W. (1979). On optimal and data-based histograms. *Biometrika, 66*, 605–610.

Scott, D. W. (1992). *Multivariate density estimation. theory, practice, and visualization.* Indianapolis, IN: Wiley.

Shapiro, S., Wilk, M., & Chen, J. H. (1968). A Comparative Study of Various Tests of Normality. *Journal of American Statistical Association, 63*, 1343-1372.

Silverman, B. W. (1985). *Density estimation for statistics and data analysis*. London, Chapman & Hall.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S-PLUS*. NY:Springer.